Journal of Biomedical Informatics 46 (2013) 894–904

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A comparative study of covariance selection models for the inference of gene regulatory networks

Patrizia F. Stifanelli [a], Teresa M. Creanza [a], Roberto Anglani [a], Vania C. Liuzzi [a], Sayan Mukherjee [c], Francesco P. Schena [b], Nicola Ancona [a,*]

[a] Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR-ISSIA, Via Amendola 122/D-O, I-70126 Bari, Italy
[b] Dipartimento Emergenza e Trapianti di Organi, DETO, Università di Bari, I-70124 Bari, Italy
[c] Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708, USA

ARTICLE INFO

Article history:
Received 21 February 2013
Accepted 8 July 2013
Available online 20 July 2013

Keywords:
Gaussian graphical models
Gene networks
Pathway analysis
Covariance selection

ABSTRACT

Motivation: The inference, or 'reverse-engineering', of gene regulatory networks from expression data and the description of the complex dependency structures among genes are open issues in modern molecular biology.

Results: In this paper we compared three regularized methods of covariance selection for the inference of gene regulatory networks, developed to circumvent the problems raising when the number of observations n is smaller than the number of genes p. The examined approaches provided three alternative estimates of the inverse covariance matrix: (a) the 'PINV' method is based on the Moore–Penrose pseudoinverse, (b) the 'RCM' method performs correlation between regression residuals and (c) 'ℓ_{2C}' method maximizes a properly regularized log-likelihood function. Our extensive simulation studies showed that ℓ_{2C} outperformed the other two methods having the most predictive partial correlation estimates and the highest values of sensitivity to infer conditional dependencies between genes even when a few number of observations was available. The application of this method for inferring gene networks of the isoprenoid biosynthesis pathways in *Arabidopsis thaliana* allowed to enlighten a negative partial correlation coefficient between the two hubs in the two isoprenoid pathways and, more importantly, provided an evidence of cross-talk between genes in the plastidial and the cytosolic pathways. When applied to gene expression data relative to a signature of *HRAS* oncogene in human cell cultures, the method revealed 9 genes (p-value < 0.0005) directly interacting with *HRAS*, sharing the same Ras-responsive binding site for the transcription factor RREB1. This result suggests that the transcriptional activation of these genes is mediated by a common transcription factor downstream of Ras signaling.

Availability: Software implementing the methods in the form of Matlab scripts are available at: http://users.ba.cnr.it/issia/iesina18/CovSelModelsCodes.zip.

1. Introduction

A challenging goal of systems biology is to provide quantitative models for the study of complex interaction patterns among genes and their products that are the result of many biological processes in the cell, such as biochemical interactions and regulatory activities [23]. Among these models, gene regulatory networks (GRNs) are essential representations for the comprehension of the development, functioning and pathology of biological organisms. Indeed, it is widely believed that the GRNs embody the comprehensive information of the mechanisms that govern the expression of the genes in the cell [28]. In particular, the GRNs inferred by genome-wide expression data depend on environmental factors, tissue type, disease-state and experimental conditions. This condition-specificity of GRNs play a major role for the study of biological processes in distinct phenotypical conditions. Indeed, under different conditions, networks exhibit different interaction patterns that can enlighten the understanding of cell development and the identification of key drivers such as disease-related genes or altered biological processes [28,51,31].

One of the simplest and most popular approaches in bioinformatics is to compute the sample Pearson correlation between every pair of genes [7]. The resulting *relevance network* considers two genes 'not-linked' in the case of *marginal* independence. This method, although useful for unveiling co-expression of genes implicated in the same biological process, has important shortcomings for the investigation of GRNs. For assessing co-expression

* Corresponding author.
E-mail address: ancona@ba.issia.cnr.it (N. Ancona).

1532-0464
http://dx.doi.org/10.1016/j.jbi.2013.07.002

between two genes, the Pearson correlation does not take into account the activities of the remaining genes in the cell. Moreover, this method does not distinguish between direct and indirect interactions, and is not able to highlight regulations by a common gene.

These drawbacks may be overcome exploiting *partial correlation*, a more sophisticated statistical model which is able to infer relations of *conditional* dependences among random variables [10,47]. In this framework, Gaussian Graphical Models (GGMs) have been exploited to study and describe dependency structures between random variables [14,26]. In our context, partial correlation assesses association between two genes by removing the effects of a set of controlling genes. Moreover, in a GGM an edge uniquely indicates a direct interaction between a gene *A* and a gene *B*, that can be interpreted biologically as one of the following mechanisms [32]:

- *A* and *B* are regulated by the same transcription factor (TF) which is not included in the network;
- *A* encodes a TF which directly regulates B;
- *A* encondes a TF which directly regulates an intermediate gene *C* which encodes a TF that in turn regulates gene *B*, and *C* is not included in the network;
- *A* encodes a protein which interacts with the TF encoded by an intermediate gene, and modifies its action on the transcription of gene *B*.

In recent years, several reverse-engineering approaches have been proposed for inferring regulatory networks from gene expression data. The nature of the data makes this problem clearly ill-posed. Indeed, the genomic data are typically characterized by a huge number p of genes and by a small number n of samples. The simplest solution proposed to overcome this problem was to reduce the numbers of genes in order to reach the $n > p$ regime [45]. Other solutions have been proposed to circumvent the problem of computing full partial correlation coefficients by using only zero and first order coefficients [48,8,19]. However, these approaches do not take into account all multigene effects on each pair of variables. More sophisticated approaches determine regularized estimates of the covariance matrix and its inverse [50,17,49]. A fundamental assumption usually adopted by these methods in $n < p$ regime is the sparsity of biological networks: only a few edges are supposed to be present in the gene regulatory networks, so that reliable estimates of the graphical model can be inferred also in small sample case [8]. A regularized GGM method based on a Stein-type shrinkage has been applied to genomic data [13] and the network selection has been based on false discovery rate multiple testing. The same procedure to select the network has been adopted, with a Moore–Penrose pseudoinverse method to obtain the precision matrix [39]. Finally, the authors in [34] suggested an attractive and simple approach based on lasso-type regression to select the non-vanishing partial correlations, paving the way to a number of analysis and novel algorithms based on lasso ℓ_1 regularizations [50,17,49,18].

To date, a comparative analysis of these methods is missing. In this work, we focus on recently proposed methods developed in the general framework of regularization and statistical learning theories which provide the state-of-art approaches for the study of ill-posed problems as the ones in which the signal is overwhelmed by the noise and the number of variables is much larger than the number of observations [46]. In particular, we focus on regularized methods for the estimation of the precision matrix in an undirected GGM. We present a comparative study of three methods in terms of AUC (area under the Receiving Operative Characteristic curve), mean square error (MSE), positive predictive values (PPV) and sensitivity (SE). The first method is based on Moore–Penrose pseudoinverse (PINV); the second one provides an estimate of the partial correlation coefficients based on Regularized Least Square regression (RCM); the third method determines an estimate of the precision matrix by maximizing a log-likelihood function properly regularized by an ℓ_2 penalty term (ℓ_{2C}). The conditional dependence between each pair of variables was assessed by using the Efron's bootstrap method [22]. Due to the lack of a perfectly known ground truth related to real biological networks [4], we measured the performance of the three methods by generating simulated data based on golden standard interaction patterns, built according to biological inspired different topologies [18,40]. We found that the ℓ_{2C} method exhibited the most predictive partial correlation estimates. More importantly, this method had the highest values of sensitivity showing its ability to infer true conditional dependencies between genes also when a few number of observations is available.

We assessed the ability of the ℓ_{2C} method to infer GRNs in two real biological contexts: the isoprenoid biosynthesis pathways in *Arabidopsis thaliana* and the *HRAS* oncogenic signature in human cell cultures. In the first case, the method enlightened known relevant pathway properties. In particular, it revealed a negative partial correlation coefficient between the two hubs in the two isoprenoid pathways. This suggests a different response of the pathways to the several tested experimental conditions and, together with the high connectivity of the two hubs, provides an evidence of cross-talk between genes in the plastidial and the cytosolic pathways. In the second case, ℓ_{2C} method highlighted 34 genes directly interacting with *HRAS*. In particular, 9 of these genes (p-value < 0.0005) shared the same Ras responsive transcription factor binding site, suggesting that their transcriptional activation is mediated by a common transcription factor downstream of Ras signaling.

2. Methods

Let $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ be a random vector distributed according a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The interaction structure among these variables can be described by means of a graph $G = (V,E)$, where V is the vertex set and E is the edge set. If vertices of V identify the random variables $X_1, \ldots, X_p$, then the edges of E represent the conditional dependence between the vertices. In other words, the absence of an edge between the ith and jth vertex means a conditional independence between the associated variables X_i and X_j. The structure of a graph is properly described by a $p \times p$ matrix, called adjacency matrix $\boldsymbol{A}$, with elements $a_{ij} = 1$ if the variables X_i and X_j (vertices) are connected by an edge and 0 otherwise.

In this study, we shall consider only undirected Gaussian graphs G with *pairwise Markov property*, such that for all $(i,j) \notin E$ one has

$$X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}} \quad i,j = 1, \ldots, p, \tag{1}$$

i.e. X_i and X_j are conditionally independent being fixed all other variables $X_{V\setminus\{i,j\}}$. Since X follows a $p-$ variate normal distribution, the condition (1) turns out to be $\rho_{ij\cdot V\setminus\{i,j\}} = 0$, where $\rho_{ij\cdot V\setminus\{i,j\}}$ is the partial correlation coefficient between the ith and jth variable, being fixed all other variables. It has been shown [26] that partial correlation matrix elements are related to the *precision matrix* (or inverse covariance matrix) $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, as:

$$\rho_{ij\cdot V\setminus\{i,j\}} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} \quad i \neq j, \tag{2}$$

where θ_{ij} are elements of $\boldsymbol{\Theta}$. In general, when the number of observations n is greater than the number of variables p, it is straightforward to evaluate θ_{ij} in Eq. (2) by inverting the sample covariance matrix. Moreover, in this case, a simple parametric test exists for

assessing the conditional independence between two variables [3]. Unfortunately, a typical genomic dataset is characterized by $n < p$, so that the sample covariance matrix becomes not invertible [11]. In the successive sections we analyze three regularized methods for estimating partial correlation matrixes and a simple non-parametric test based on Efron's bootstrap method to use when $n < p$ for assessing conditional independence [22].

2.1. Partial correlation matrix estimation

For describing the three methods that we have analyzed, let us consider the $n \times p$ data matrix

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$$

where each column $\{\mathbf{X}_i\}_{i=1,\ldots,p} \in \mathbb{R}^n$ is a $n-$dimensional vector, with $n < p$. Let $\mathbf{S}$ be the sample estimate of the covariance matrix $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Theta}}$ be the estimate of inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$.

2.1.1. Pseudoinverse method (PINV)

The estimated precision matrix $\hat{\boldsymbol{\Theta}}$ can be obtained as pseudoinverse of $\mathbf{S}$, by using the Singular Value Decomposition (SVD). Indeed, since $\mathbf{S}$ is a real and symmetric matrix, then its singular value decomposition reduces to $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ where $\mathbf{U}$ is a $p \times p$ unitary matrix whose columns are the eigenvectors of $\mathbf{S}$ and $\mathbf{U}^\top$ is the transpose of $\mathbf{U}$; $\boldsymbol{\Lambda}$ is $p \times p$ diagonal matrix whose entries are the non-negative eigenvalues of $\mathbf{S}$. Then, the pseudoinverse of $\mathbf{S}$ is $\mathbf{S}^+ = \mathbf{U}\boldsymbol{\Lambda}^+\mathbf{U}^\top$, where $\boldsymbol{\Lambda}^+$ is obtained by replacing each positive diagonal element of $\boldsymbol{\Lambda}$ with its reciprocal.

To improve the estimate of the partial correlation coefficients, we evaluated a bootstrap version of $\hat{\boldsymbol{\Theta}}$ [39]. In particular, we generated B bootstrap replications $\mathbf{X}^b$ of the sample with $b = 1, \ldots, B$ obtained by random sampling with replacement the raws of $\mathbf{X}$. For each replication, we evaluated the bootstrap replication $\mathbf{S}^b$ of $\mathbf{S}$ and used these estimates for obtaining the bootstrap mean $\mathbf{S}_B = \frac{1}{B}\sum_{b=1}^{B}\mathbf{S}^b$. Then the bootstrap estimate of $\hat{\Theta}$ was obtained as $\hat{\Theta}_B = \mathbf{S}_B^+$.

Finally, we estimated the partial correlation matrix as

$$\hat{\boldsymbol{\rho}} = -\frac{\hat{\boldsymbol{\Theta}}_B}{\sqrt{\mathrm{diag}\hat{\boldsymbol{\Theta}}_B \mathrm{diag}\hat{\boldsymbol{\Theta}}_B^\top}}. \tag{3}$$

2.1.2. Covariance-regularized method (ℓ_{2C})

Let us consider the loss function [3]

$$L_l(\mathbf{S}, \boldsymbol{\Theta}) = \mathrm{Tr}(\mathbf{S}\boldsymbol{\Theta}) - \log\det(\mathbf{S}\boldsymbol{\Theta}) - p \tag{4}$$

that vanishes when $\mathbf{S}\boldsymbol{\Theta} = \mathbf{I}$ and is positive when $\mathbf{S}\boldsymbol{\Theta} \neq \mathbf{I}$. Since we are dealing with the case of $p > n$, an estimate of $\boldsymbol{\Theta}$ could be obtained minimizing with respect to $\boldsymbol{\Theta}$ the ℓ_2-penalized loss function:

$$L_p(\mathbf{S}, \boldsymbol{\Theta}, \lambda) = L_l(\mathbf{S}, \boldsymbol{\Theta}) + J(\lambda, \boldsymbol{\Theta}), \tag{5}$$

where

$$J(\lambda, \boldsymbol{\Theta}) = \lambda\|\boldsymbol{\Theta}\|_F^2 \tag{6}$$

with $\lambda > 0$ and $\|\boldsymbol{\Theta}\|_F^2 = \mathrm{tr}(\boldsymbol{\Theta}^\top\boldsymbol{\Theta})$ is the Frobenius norm of $\boldsymbol{\Theta}$.

Note that, the minimization of $L_p(\mathbf{S}, \boldsymbol{\Theta}, \lambda)$ with respect to $\boldsymbol{\Theta}$ is equivalent to the maximization of the penalized log-likelihood [49]:

$$\log\det(\boldsymbol{\Theta}) - \mathrm{Tr}(\mathbf{S}\boldsymbol{\Theta}) - \lambda\|\boldsymbol{\Theta}\|_F^2. \tag{7}$$

Differentiating with respect to $\boldsymbol{\Theta}$ means to solve the following equation

$$\hat{\boldsymbol{\Theta}}^{-1} - 2\lambda\hat{\boldsymbol{\Theta}} = \mathbf{S}. \tag{8}$$

Consequently, the problem turns out to be an eigenvalue problem. Indeed, if θ_i are eigenvalues of $\hat{\boldsymbol{\Theta}}$ with eigenvectors $\mathbf{u}_i$,

$$\hat{\boldsymbol{\Theta}}\mathbf{u}_i = \theta_i\mathbf{u}_i \tag{9}$$

then s_i are the eigenvalues of $\mathbf{S}$ with the same eigenvectors, and the relation between θ_i and s_i is $\theta_i^{-1} - 2\lambda\theta_i = s_i$. Therefore the eigenvalues θ_i of $\hat{\boldsymbol{\Theta}}$ can be evaluated as function of the eigenvalues s_i of $\mathbf{S}$:

$$\theta_i^{\pm} = -\frac{s_i}{4\lambda} \pm \frac{\sqrt{s_i^2 + 8\lambda}}{4\lambda}. \tag{10}$$

Since precision matrix must be positive definite, the correct value of θ_i is θ_i^+. Then, for the spectral theorem, $\hat{\boldsymbol{\Theta}}$ is given by

$$\hat{\boldsymbol{\Theta}} = \sum_{i=1}^{\ell}\theta_i^+\mathbf{u}_i\mathbf{u}_i^\top. \tag{11}$$

The regularization parameter λ was selected by using the cross validation procedure. In particular, for each value of λ in a suitable range, we carried out 20 random splits of the dataset in training $\mathbf{X}_t$ and validation $\mathbf{X}_v$ sets and evaluated the corresponding sample covariance matrices $\mathbf{S}_t$ and $\mathbf{S}_v$. Consequently, we estimated $\boldsymbol{\Theta}_t^\lambda$ by minimizing the penalized loss function in Eq. (5) and evaluated the loss function in Eq. (4), averaged over the 20 splits, $\langle L_l(\mathbf{S}_v, \boldsymbol{\Theta}_t^\lambda)\rangle$. The selected λ value was

$$\lambda^* = \arg\ \min_\lambda\langle L_l(\mathbf{S}_v, \boldsymbol{\Theta}_t^\lambda)\rangle. \tag{12}$$

This procedure selected the lambda minimizing the distance between the empirical inverse precision matrix computed on the training set $\mathbf{X}_t$ and the sample covariance matrix computed on the validation set $\mathbf{X}_v$.

2.1.3. Residual correlation method (RCM)

Let us consider a linear regression model for the variables $\mathbf{X}_i$ and $\mathbf{X}_j$ given all the $p-2$ remaining variables:

$$\mathbf{X}_i = \mathbf{X}_{\setminus i\setminus j}\boldsymbol{\beta}_i \qquad \mathbf{X}_j = \mathbf{X}_{\setminus i\setminus j}\boldsymbol{\beta}_j \tag{13}$$

where $\boldsymbol{\beta}_i \in \mathbb{R}^{p-2}$ is the regression coefficient vector referred to the ith gene; $\mathbf{X}_i$ is the ith column of the matrix $\mathbf{X}$ and $\mathbf{X}_{\setminus i\setminus j}$ is $\mathbf{X}$ without the ith and jth column. Note that the bias term is implicitly present in our model. This is done by including a component constant and equal to one to the input vectors. The Regularized Least Square (RLS) [20,2] method evaluates the regression models (13) by solving

$$\min_{\boldsymbol{\beta}_i \in \mathbb{R}^{p-2}} \frac{1}{n}\|\mathbf{X}_i - \mathbf{X}_{\setminus i\setminus j}\boldsymbol{\beta}_i\|_2^2 + \mu\|\boldsymbol{\beta}_i\|_2^2. \tag{14}$$

where $\mu > 0$ is the regularization parameter. If $\hat{\mathbf{X}}_i$ and $\hat{\mathbf{X}}_j$ are the RLS estimates of $\mathbf{X}_i$ and $\mathbf{X}_j$, one can evaluate the residual vectors $\mathbf{r}_i = \hat{\mathbf{X}}_i - \mathbf{X}_i$ and $\mathbf{r}_j = \hat{\mathbf{X}}_j - \mathbf{X}_j$. This allows to evaluate the partial correlation coefficient $\hat{\rho}_{ij|p-2}$ between the ith and jth variable being fixed all other $(p-2)$ variables as the Pearson correlation $r_{r_i r_j}$ between the residuals, i.e.

$$\hat{\rho}_{ij|(p-2)} = r_{r_i r_j} = \frac{\mathrm{cov}(\mathbf{r}_i, \mathbf{r}_j)}{\sqrt{\mathrm{var}(\mathbf{r}_i)\mathrm{var}(\mathbf{r}_j)}}. \tag{15}$$

Finally, the μ parameter was chosen by minimizing the Leave-One-Out cross validation errors [30].

2.2. Non-parametric test for conditional independence

The conditional independence structure among variables and then the estimated graph $\hat{\mathbf{A}}$ was inferred calculating a 95% confidence interval for each entry ρ_{ij}, by using Efron's bootstrap method [22]. The graph selection procedure is as follows:

(I) Build $B = 100$ bootstrap replications by drawing randomly with replacement n rows from $\mathbf{X}$.
(II) Evaluate $\hat{\boldsymbol{\rho}}^b$ for each $b = 1, \ldots, B$.
(III) Rank the bootstrap replications $\hat{\rho}_{ij}^{(1)} \leqslant \hat{\rho}_{ij}^{(2)} \leqslant \ldots \leqslant \hat{\rho}_{ij}^{(B)}$.
(IV) Compute the 95% confidence interval $(\rho_{ij}^L, \rho_{ij}^U)$ where $\rho_{ij}^L = \hat{\rho}_{ij}^{(k)}, \rho_{ij}^U = \hat{\rho}_{ij}^{(B+1-k)}, k = B\frac{\alpha}{2}$ and $\alpha = 0.05$.
(V) $\hat{a}_{ij} = 1$ if $0 \notin (\rho_{ij}^L, \rho_{ij}^U), \hat{a}_{ij} = 0$ otherwise.

2.3. Simulation study

We conducted an extensive simulation study for analyzing the performances of the three methods in terms of their ability to infer conditional dependency structures among variables in the case of $n \ll p$. In a nutshell, the simulation study was composed of three main steps: (a) building a graph with a well defined and biologically inspired structure and the corresponding precision matrix; (b) generating observations of the p random variables according to the dependency structure defined by the graph; and (c) comparing the inferred graph with the one used for generating the data in terms of estimated partial correlations and predicted edges.

2.3.1. Graph building

We built gold standard graphs G_{GoS} by defining adjacency matrices $\mathbf{A}_{\text{GoS}}$ according to three different kinds of patterns [18,36]:

Random: $\mathbf{A}_{\text{GoS}}$ is randomly generated by inserting approximately p nonzero entries;
Hubs: the rows/columns of $\mathbf{A}_{\text{GoS}}$ are partitioned into K disjoint groups of q variables. Each group consists of five hubs with high degree, and the other $q - 5$ nodes with lower degrees. This setting is designed to simulate scale-free-like gene regulatory networks, which tipically contain a few hub genes plus many other nodes with only a few connections.
Clique: the rows/columns of $\mathbf{A}_{\text{GoS}}$ are partitioned into K disjoint groups of q variables fully connected, i. e. each group is a clique.

The density d of $\mathbf{A}_{\text{GoS}}$ is defined as:

$$d = \frac{E}{p(p-1)/2} \tag{16}$$

where E is the number of edges of the graph and $p(p-1)/2$ is the size of a complete graph with p nodes.

2.3.2. Data set generation

Given a gold standard graph defined by an adjacency matrix $\mathbf{A}_{\text{GoS}}$, there exists a family of precision matrices $\boldsymbol{\Theta}_{\text{GoS}}$, or equivalently of partial correlation matrices $\boldsymbol{\rho}_{\text{GoS}}$, associated to the graph constituted by all the real positive definite $p \times p$ symmetric matrices having zeros in the same positions as $\mathbf{A}_{\text{GoS}}$ [8]. In order to simplify the simulation study, we decided to build partial correlation matrices having constant non-zero entries. In other words, the partial correlation between any pair of conditionally dependent variables was constant all over the graph. To this end, for a given adjacency matrix $\mathbf{A}_{\text{GoS}}$, the associated concentration matrix $\boldsymbol{\Theta}_{\text{GoS}}$ was built as:

$$\Theta_{\text{GoS}} = \frac{1}{m+\epsilon}\mathbf{A}_{\text{GoS}} + \mathbf{I_p} \tag{17}$$

where $m = \max_i \sum_j a_{ij}, \epsilon > 0$ and $\mathbf{I}_p$ is the p order identity matrix. Since $\mathbf{A}_{\text{GoS}}$ is a symmetric matrix having diag $(\mathbf{A}_{\text{GoS}}) = \mathbf{0}$ by definition, then diag $(\boldsymbol{\Theta}_{\text{GoS}}) = \mathbf{I}_p$. As a consequence, $\boldsymbol{\rho}_{\text{GoS}} = -\boldsymbol{\Theta}_{\text{GoS}}$ (see Eq. 2). Moreover, the precision matrix defined in Eq. (17) is strictly diagonally dominant and then is positive definite. Finally, the data set $\mathbf{X}$ was generated by sampling n times a p-variate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{GoS}})$, with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\text{GoS}}$, where $\boldsymbol{\Sigma}_{\text{GoS}} = \boldsymbol{\Theta}_{\text{GoS}}^{-1}$.

2.3.3. Performance measures

The performances of the three methods were assessed by using different criteria. The first one aimed to quantify the accuracy of a method in estimating the partial correlation values. To this end, we evaluated the mean square error (MSE) between the estimated $\hat{\boldsymbol{\rho}}$ and gold standard $\boldsymbol{\rho}_{\text{GoS}}$ partial correlation matrices:

$$\text{MSE} = \sqrt{\frac{2}{p(p-1)}\sum_i\sum_{j>i}(\hat{\rho}_{ij} - \rho_{ij}^{\text{GoS}})^2}. \tag{18}$$

The second criterion aimed to assess a method in terms of prediction accuracy of the entries of $\mathbf{A}_{\text{GoS}}$ by using the estimated partial correlation values $\hat{\boldsymbol{\rho}}$ as predictor variable. In particular, we wanted to quantify the prediction error of a binary classifier which predicts the entries of the gold standard adjacency matrix $\mathbf{A}_{\text{GoS}}$ as edge ($a_{ij}^{\text{GoS}} = 1$) or non edge ($a_{ij}^{\text{GoS}} = 0$) based upon $\hat{\rho}_{ij}$. To this end, we evaluated the Area Under the ROC Curve (AUC) which assesses the performances of a binary classifier as its discrimination threshold is varied. AUC is equal to the probability that a classifier will rank randomly a chosen positive instance higher than a randomly chosen negative one [15]. We chose AUC as measure of performance because it is not influenced by the prevalence of a class and it is independent from the selection rule used to infer the graph. The AUCs evaluated for the three methods were compared with the AUC measured for a random algorithm which assigns randomly and with the same probability 0 or 1 independently of the estimated partial correlation values [4]. The third criterion aimed to evaluate a method by comparing the estimated graph $\hat{\mathbf{A}}$, inferred by using $\hat{\boldsymbol{\rho}}$, with the gold standard graph given by $\mathbf{A}_{\text{GoS}}$. Positive Predicted Values (PPV) and sensitivity (SE) were evaluated for comparing $\hat{\mathbf{A}}$ and $\mathbf{A}_{\text{GoS}}$. The last criteria consisted in measuring the computational time required for estimating the $\hat{\boldsymbol{\rho}}$ matrix.

3. Results

3.1. Results on simulated data

The performances of the three methods were analyzed keeping constant the number of variables to $p = 400$, while varying the number of samples n in the range [30] with step 30. For each value of n, the results were averaged over 20 repetitions. Moreover, for each value of n and for each generated graph structure (random, hubs, clique), all the methods were applied on the same simulated data sets.

3.1.1. MSE analysis

The methods exhibited very different accuracies in the estimate of the true partial correlation values. The comparison was carried out keeping constant the graph density to $d = 0.01$ and evaluating MSE both globally on the whole partial correlation matrix $\boldsymbol{\rho}_{\text{GoS}}$, and evaluating MSE limitedly to the non null entries of $\boldsymbol{\rho}_{\text{GoS}}$ corresponding to the true edges of the gold standard graph (see Fig. 1). We denote these two errors as MSE and edge-MSE. This choice was due to the fact that the true partial correlation matrices used in our simulations were generally sparse having only a few non-null entries. For small values of n, RCM method exhibited poor accuracy, showing the highest MSE and edge-MSE errors. The behavior of the two errors was comparable indicating that the method behaves uniformly on the whole graph. Although insufficient in accuracy, the method had performances invariant with respect to the graph topology and its accuracy increased with n.

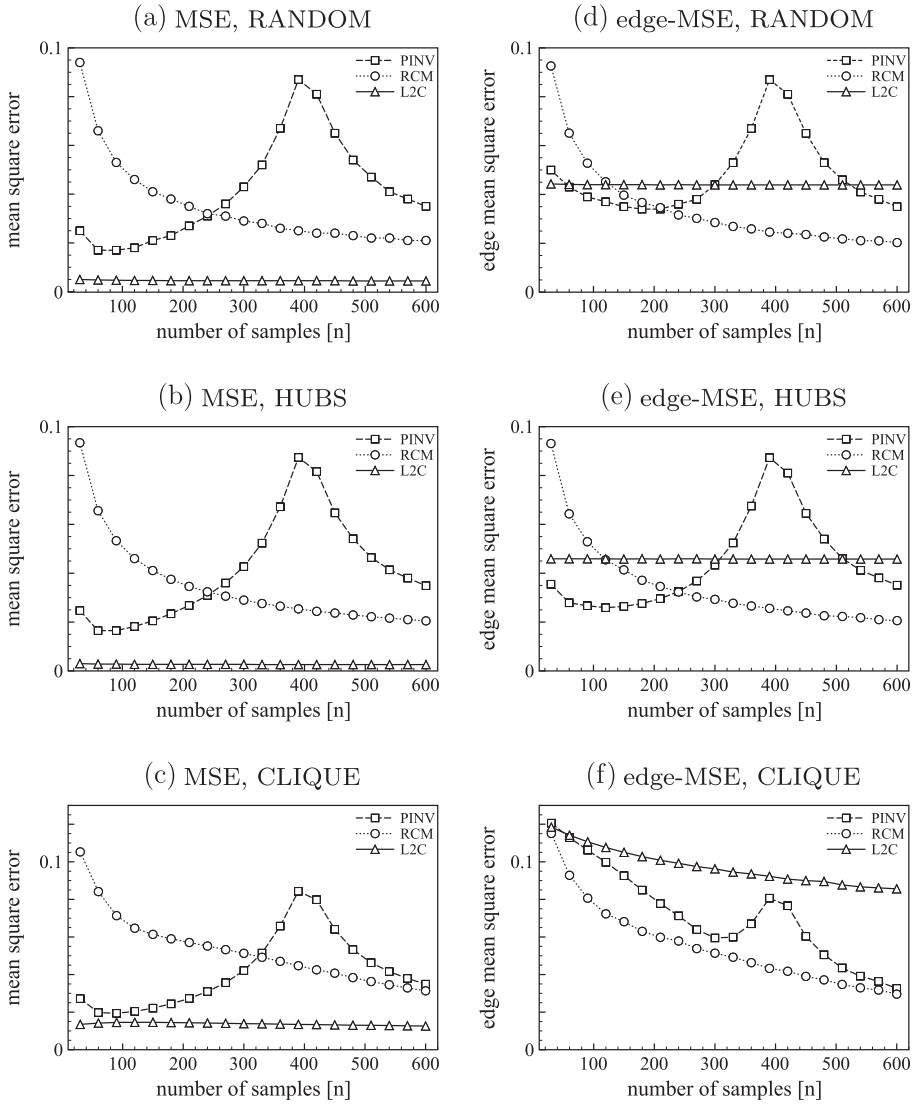

Fig. 1. Comparison of the methods by varying the graph topology and n, with a graph density of $d = 0.01$. The left panel (pictures (a)–(c)) shows the mean square error (MSE) evaluated on the whole $\boldsymbol{\rho}_{GoS}$ matrix; the right panel (pictures (d)–(f)) shows the mean square error evaluated limitedly to the non null entries of $\boldsymbol{\rho}_{GoS}$ (edge-MSE).

A better accuracy was reached by PINV and ℓ_{2C} methods, specially for small values of n. The performances of PINV method were not influenced by the graph topology and the MSE and edge-MSE errors remained comparable in all the simulations. Nevertheless, we observed a "resonance effect" associated to this method for $n \approx p$, confirming a behavior well known in literature [37,39]. Increasing the number of observations n up to p the performance of the method got worse. For values of n larger than p, the error associated to PINV decreased by increasing n.

Contrarily to RCM and PINV, the accuracy of the ℓ_{2C} method was found different when evaluated on the whole graph or limitedly to the edges only (see Fig. 1). In particular, we measured MSE ≈ 0.01 and edge-MSE ≈ 0.04 for both random and hubs graphs, independently to n. A larger difference was found in clique graph topology. This discrepancy in accuracy is due to the regularization term introduced in the penalized loss function (see Eq. (7)). Indeed, ℓ_{2C} method selects precision matrices with small Frobenius norm and as a consequence, underestimates the partial correlation values. The important aspect of this method is that its accuracy is poorly influenced by the number of available observations and by the graph intrinsic topology.

3.1.2. AUC analysis

The analysis of the performances in terms of AUC highlighted other interesting properties of the methods. While the MSE analysis evaluated the methods by comparing the *true* and the *estimated* partial correlation values, the AUC analysis assessed the methods by comparing the *accuracy* in the prediction of the gold standard graph by using the partial correlation estimates produced by the methods. The important aspect to underline is that AUC analysis is independent of the prediction rule and highlights properties of the variable used as predictor. The Fig. 2 depicts the behaviors of AUC evaluated for the three methods by varying n, for $p = 400$ and $d = 0.01$. All the three methods provided satisfactory results also for very small values of n because they outperformed significantly the performances of a random algorithm (see Supplemental Information). For example, in the case of random graph topology, ℓ_{2C} and RCM exhibited AUC = 0.64 with 95% CI [0.59, 0.67] for $n = 30$ observations; PINV showed AUC = 0.61 with 95% CI [0.56, 0.65] for $n = 60$ observations. In general, our simulations showed that ℓ_{2C} and RCM provided partial correlation estimates with a prediction accuracy higher than PINV. Moreover, for these two methods, it was sufficient to exploit a number of observations greater than the 15% of the number of the variables for having an

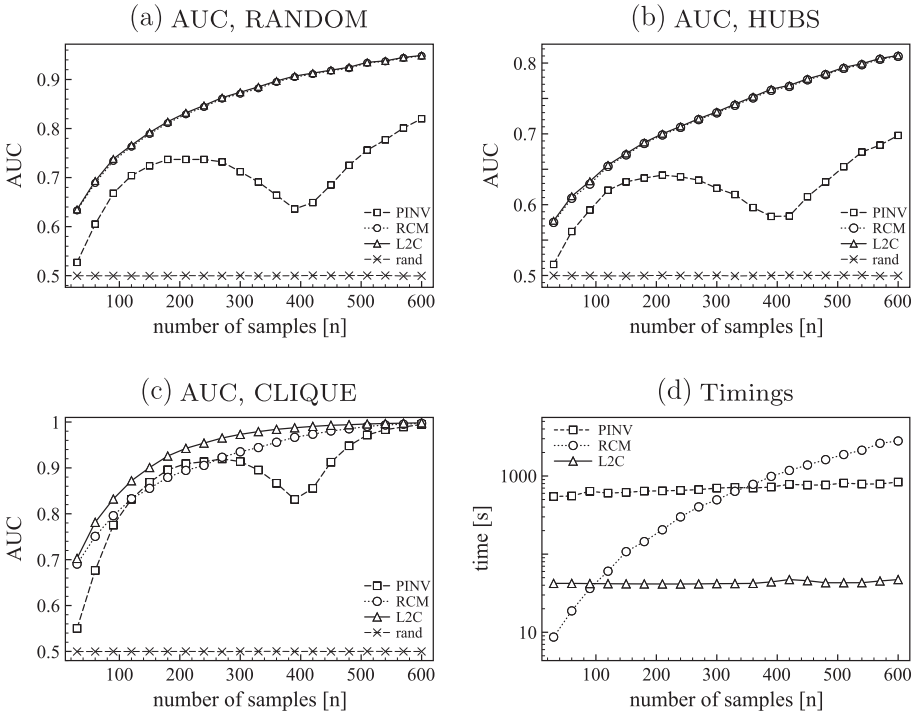

Fig. 2. Comparison of the methods in terms of AUC for (a) random, (b) hubs, (c) clique graph topologies by varying n, with a density of $d = 0.01$. The '-x-' line style depicts the AUC of a random classifier. (d) Comparison of the methods in terms of the computational time required for estimating the $\hat{\boldsymbol{\rho}}$ matrix. The timings are in seconds.

AUC value greater than 0.7. The accuracy of the partial correlation values estimated by ℓ_{2C} and RCM methods improved by increasing the number of observations. On the contrary, the estimates produced by PINV method suffered for $n \approx p$ of the same instability found in MSE analysis.

Another interesting property highlighted by the AUC analysis was the dependency of the accuracy on graph topology. In fact, as Fig. 2(a)–(c) show, the best accuracies were obtained for clique graphs and the worse for hub graphs. The reason of these different performances, obtained keeping constant the density d in all the simulations, resides in the *degree* distribution of the graph. With the degree term we mean the number of nodes connected to a given node, or equivalently, the number of variables conditionally dependent on a given variable. In fact, in the case of clique topology, the graph was composed of $K = 80$ disjoint cliques, each composed of $q = 5$ variables. In this case, the number of variables conditionally dependent on a given variable had a constant value of $q - 1$ and this number was also the maximum number of variables conditionally dependent. In the case of random graphs, the number of edges connected to a given node had a binomial distribution $B(p - 1, d)$. So, with a density of $d = 0.01$ and $p = 400$, we had a mean degree of $pd = 4$, with a range of $[0, 11]$. This is equivalent to saying that 11 was the maximum number of variables conditionally dependent on a given variable. Finally, in the case of hub topology, the graph was composed of $K = 10$ disjoint groups each composed of $q = 40$ variables, giving a mean degree equal to 4 with a range of $[0, 19]$.

The dependency of the AUC on density and graph topology was confirmed by two different simulations (Fig. 3). In the first one, limitedly to the random graph topology, each method was analyzed by varying the density of the graph. In the second one, keeping the density fixed to $d = 0.01$, each method was analyzed by varying the graph topology. As the Fig. 3(a), (c) and (e) show, the lower the density, the more accurate the estimates for all the methods. Moreover, as the Fig. 3(b), (d) and (f) show, the accuracy of the estimates decreased going from clique to hub graphs for all the methods consistently.

3.1.3. PPV and sensitivity analysis

The estimate of the partial correlation values is only the first step of any graph inference procedure and the accuracy of the estimates provided by the three methods was assessed in the two previous sections. With the sensitivity analysis addressed in this section, we assessed the whole graph inference procedure described in the Methods section, which exploited the estimates provided by the methods for inferring a graph. As performance measure we used PPV and sensitivity evaluated comparing the gold standard graph with the one inferred by the procedure. For each value of n, the values were averaged over 20 repetitions and evaluated for hub graph topology. Fig. 4 depicts the behaviors of PPV and sensitivity as a function of n. PINV was the only method that, for small values of n, had PPV and sensitivity values smaller than those provided by the random algorithm. RCM and ℓ_{2C}, on the contrary, showed PPV and sensitivity values greater than the random algorithm, even for a small number of observations. In particular, ℓ_{2C} method exhibited the best performances in terms of sensitivity, consistently for all the values of n, indicating that this method was able to infer true conditional dependences between variables also when a few number of observations is available. Note that the PPV and sensitivity values shown in Fig. 4 have not to be considered as absolute indicators of the performances of a method. On the contrary, they have to be considered as indicative of the relative performances of a method with respect to another one, because assessed for a specific experimental condition, equal for all the methods. More importantly, these values have to be significantly different from those obtained by random algorithms. Moreover, the values of PPV and sensitivity of a method heavily depend on the strength of the partial correlation existing among the variables. The partial correlation values used in our simulations

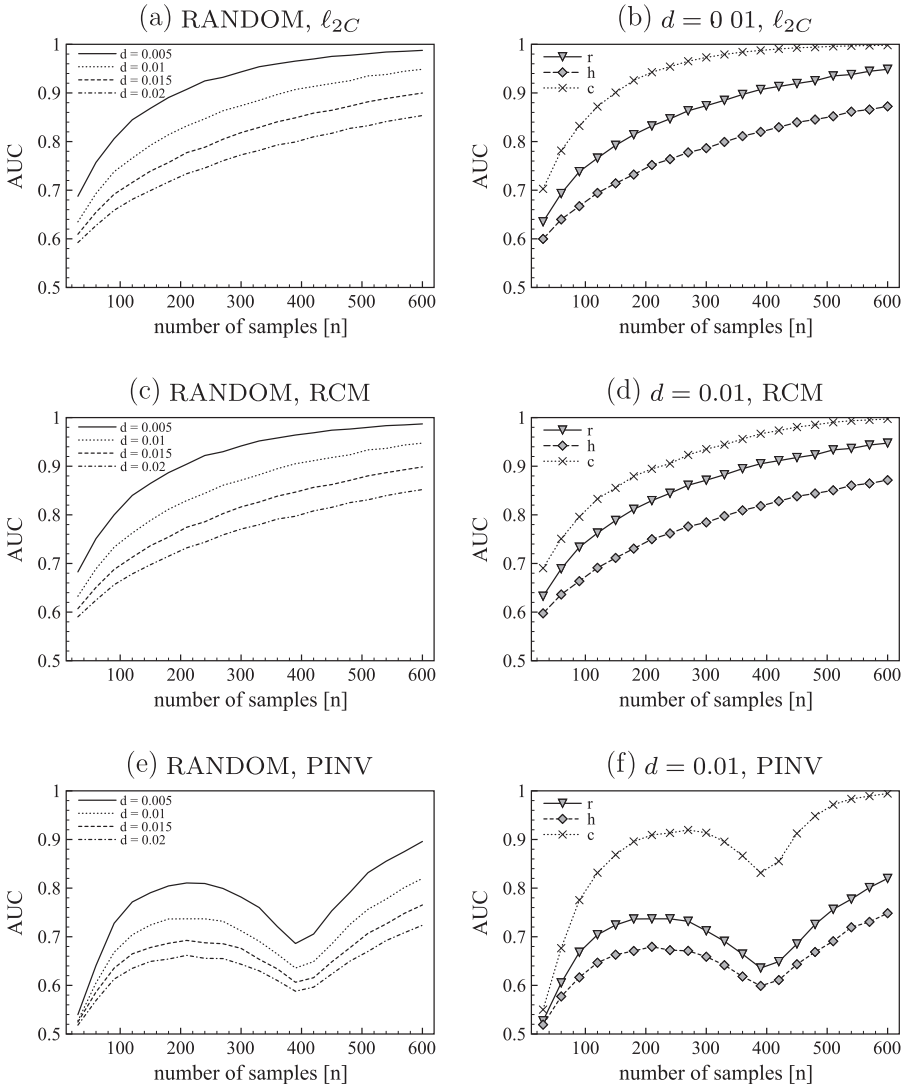

Fig. 3. Performances of a method in terms of AUC. *Left panel*: constant graph topology and variable *d*. *Right panel*: constant *d* and variable graph topology.

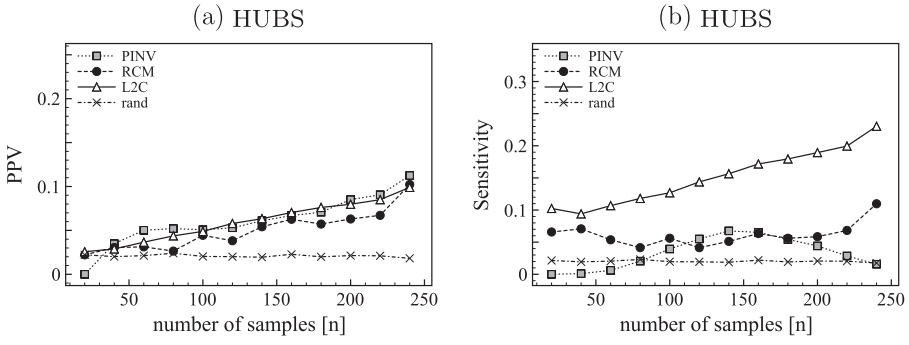

Fig. 4. PPV and Sensitivity of the methods as a function of *n*.

were close to zero (see Eq. 17), so generating critical experimental conditions for the methods.

In the light of the simulation results described so far, and considering the computational time required by the three methods (see Fig. 2(d)), we choose ℓ_{2C} as the method to apply for inferring biological networks. This choice was motivated by the AUC behavior which outperformed the other two methods in all the considered graph topologies and, mainly, by the sensitivity of the method.

3.2. Application to gene expression data

We applied ℓ_{2C} method for the inference of gene regulatory networks from DNA microarray data in two different contexts. The first concerned the cross-talk between the two isoprenoid pathways of the model system *A. thaliana*. In this case we applied the method to a well studied benchmark data set for the inference of gene networks [48,19] and compared our findings with the ones reported in literature. The second concerned the investigation of

the genes interacting with the oncogene *HRAS*. In this case we applied the method to a gene expression data set with strong a priori biological knowledge [6]. In fact these data were used for inferring a signature of *HRAS* through its in vitro overexpression.

3.2.1. Isoprenoid pathways in A. thaliana

The isoprenoids are a large class of organic compounds derived from isoprene. They play various important roles in plants as: quinones in electron transport chains, structural components of membranes, photosynthetic pigments, hormones, defense compounds, attractants for pollinators and in subcellular targeting and regulation [21]. Isoprenoids are synthesized through condensation of the five-carbon intermediates isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) [24]. In higher plants IPP and DMAPP are synthesized through two different routes that take place in two distinct cellular compartments. The cytosolic pathway, also called MVA pathway, starts from acetyl-CoA and moves through the intermediate mevalonate (MVA), providing the precursors for sterols, ubiquinone and sesquiterpenes [12]. An alternative pathway, called non-mevalonate pathway or MEP/DOXP pathway, is located in the chloroplast. It implicates the condensation of pyruvate and glyceraldehyde-3-phosphate via 1-deoxy-D-xylulose 5-phosphate (DOXP) and 2-C-methyl-D-erythritol 4-phosphate (MEP) and is used for the synthesis of isoprene, carotenoids, abscisic acid, and the side chains of chlorophylls and plastoquinone [29]. Although this subcellular compartmentation allows both pathways to work independently, there are several evidences that they can interact in some conditions [25,38,5]. Inhibition of the cytosolic MVA pathway in *A. thaliana* leds to an increase of levels of carotenoids and chlorophylls, demonstrating that the decreased working of MVA pathway can be in part compensated for by the MEP pathway. Inversely, inhibition of the MEP pathway in seedlings causes the reduction of carotenoids and chlorophylls levels, indicating a predominantly unidirectional transport of isoprenoid intermediates from the chloroplast to the cytosol. In order to investigate whether the transcriptional regulation is at the basis of the crosstalk between the cytosolic and the plastidial pathways, Laule et al. [25] studied this interaction by identifying the genes with expression levels changed as a response to the inhibition. They have shown that the inhibitor mediated changes in metabolite levels are not reflected in changes in gene expression levels, suggesting that alterations in the flux through the cytosolic and plastidial pathways of isoprenoid metabolism are not transcriptionally regulated. In order to clarify the interaction between the two pathways at the transcriptional level, Wille and Buhlmann [48] have explored the structural relationship between genes on the basis of their expression levels under different experimental conditions. This study aimed to infer the regulatory network of the genes in the isoprenoid pathways by incorporating the expression levels of 795 genes from other 56 metabolic pathways. Moving beyond the one-gene approach, the authors have found various connections between genes in the two different pathways, suggesting the existence of a crosstalk at the transcriptional level.

We applied the ℓ_{2C} method to the publicly available data set from [48]. The data consisted of expression measurements for 39 genes in the isoprenoid pathways and 795 in other 56 pathways assayed on 118 Affymetrix GeneChip microarrays. Among the 39 genes in the isoprenoid pathways, 15 are assigned to the cytosolic pathway, 19 to the plastidal pathway and 5 encode mitochondrial proteins involved in isoprenoid synthesis. We were interested in the construction of a gene network related to the two isoprenoid pathways considering also the effects of genes in the other pathways. To this end, we built 1000 bootstrap replications of the data set and used 95% confidence interval for inferring the network. The Fig. 5 depicts the inferred network with 44 edges. For each pathway we found a module with strongly interconnected and positively correlated genes. This enlightens the reliability of our method since genes within the same pathway are potentially jointly regulated [42]. Furthermore, we identified two strong candidate genes for the cross-talk between both pathways: HMGS and HDS. HMGS represents the hub of the cytosolic module since it is positively correlated to five genes of the same pathway: DPPS1, MDPC1, AACT2, HMGR2 and MK. It encodes a protein with hydroxymethylglutaryl-CoA synthase activity that catalyses the second step of the MVA pathway. HDS represents the hub of the plastidial module since it is positively correlated to five genes of the same

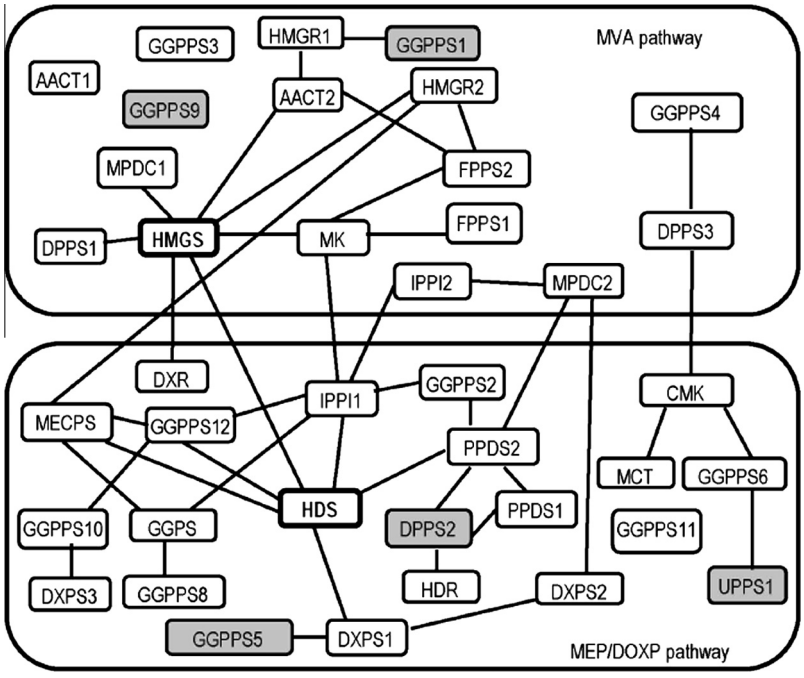

Fig. 5. Biological network of the isoprenoid pathways inferred by using ℓ_{2C}. *Upper part*: Genes of MVA pathway. *Lower part*: Genes of MEP/DOXP pathway. Grey boxes refer to mithochondrial genes; HMGS and HDS represent the candidate hubs of the two modules.

pathway: DXPS1, MECPS, GGPPS12, IPPI1 and PPDS2. It encodes a chloroplast-localized hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase and catalyses the penultimate step of the biosynthesis of IPP and DMAPP via the MEP/DOXP pathway. The negative correlation between HMGS and HDS suggests that they respond differently to the tested experimental conditions. This, together with the high connectivity of the two hubs, provides an evidence of cross-talk between proteins in the plastidial and the cytosolic pathways. Other negative correlations between the two pathways are represented by the edges HMGR2|MECPS, MPDC2|PPDS2 and MPDC2|DXPS2. Interestingly, the plastidial gene IPPI1 is found to be positively correlated to the module of connected genes in the MVA pathway (IPPI1|MK, IPP1|IPPI2). This evidence confirms the results of [19] where they guessed that the enzyme IPPI1 controls the steady-state levels of IPP and DMAPP in the chloroplast, when a high level of transfer of intermediates between the two cell compartments takes place. Moreover, our study showed three candidate mitochondrial genes for the cross-talk (DPPS2, GGPPS5 and UPPS1) which are in the plastidial module. Finally, it is interesting to note that the method used in [48] and in [19] included more cross-links between the two pathways with respect to the ℓ_{2C} method. Although it is known the existence of cross-links between the two pathways, we believe that these interactions should not be so numerous, as genes of the two pathways belong to two different cell compartments. A possible explanation of such a difference is that [48,19] constructed a network based on first-order conditional dependencies that are not able to capture all multi-gene effects on a given pair of genes.

3.2.2. Interacting genes in HRAS signature

Ras genes represent a GTPase superfamily composed by more than 150 distinct cellular members, among which the most representatives are *HRAS*, *NRAS* and *KRAS*. Up to 30% of all screened human tumors are found to carry some mutations in any of these genes. Ras signal transduction proceeds through activation of some signal transduction cascades, such that of Mitogen-Activated Protein Kinases (MAPKs), and culminates in the modulation of transcription of specific genes involved in many physiological processes including cell cycle progression, growth, migration, cytoskeletal changes, apoptosis, and senescence. The cross-talk among this plethora of actors creates a molecular network whose balance is crucial to determine normal cellular responses. Indeed, alterations of Ras signaling could break this balance and induce the onset of cancer and for this reason the inference and the analysis of Ras network is of fundamental importance [16].

In this context, we applied the ℓ_{2C} method for inferring genes directly interacting with *HRAS*. To this end, we used a data set with a controlled genetic perturbation of *HRAS* used to generate its oncogenic signature [6]. Such a signature was identified by infection of human primary mammary epithelial cell cultures (HMECs) with adenoviruses expressing activated *HRAS*. The signature was composed of those 276 genes for which the expression levels were mostly correlated with the classification of HMEC samples into *HRAS*-activated versus wild-type. The resulting data set used in our experiment was composed of 276 genes assayed in 10 samples relative to *HRAS*-activated and 10 samples relative to wild-type *HRAS*. Indeed, we considered that the RAS signature retrieved by

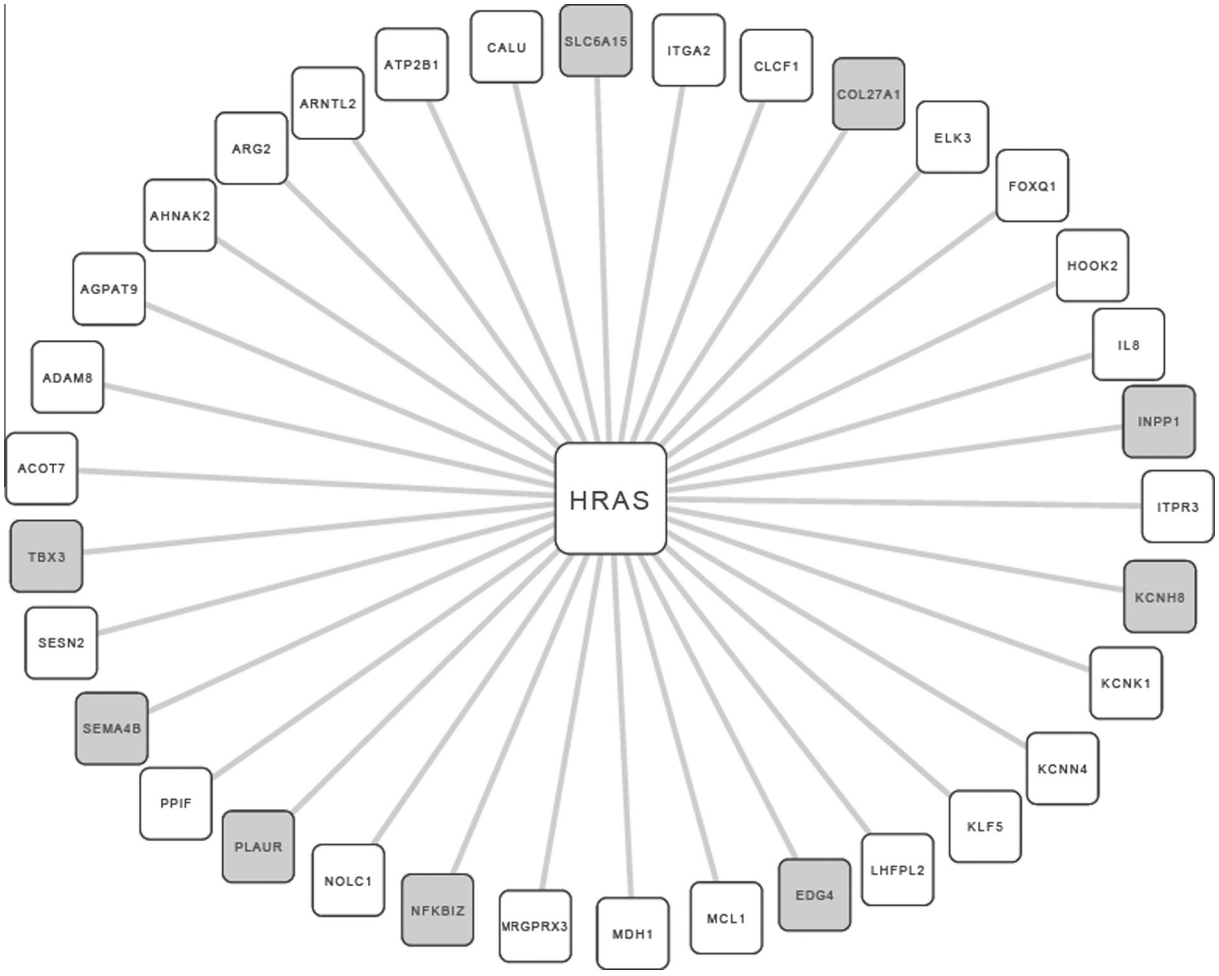

Fig. 6. Biological network of the 34 genes interacting with *HRAS* inferred using ℓ_{2C} method. The gray shaded boxes indicate the genes sharing the RREB1 consensus binding site.

Bild includes direct and indirect connections of the H-RAS gene with the others and, consequently, we applied L2C method in order to select only the direct interactions. Moreover, the conditional dependences were evaluated conditioning over only the genes in the signature because the study of Bild suggests that the interactions between the signature and the other genes on chip can be neglected.

The selection of the graph was performed by building 1000 bootstrap replications of the data and computing the 99,5% confidence interval of the statistics. The resulting network was composed of 2875 edges (see Additional file 4), where *HRAS* had 34 direct connections (Fig. 6).

We analyzed the 34 *HRAS* interacting genes with the TRANSFAC component of GATHER [9] to assess the significance of the presence of common potential transcription factor binding sites within their promoters. A very interesting finding was that the list of these 34 genes was enriched of the RREB1 (Ras responsive element binding protein 1) module with a p-value < 0.0005. In fact, 9 of them (see Fig. 6) presented the RREB1 consensus binding site. On the contrary, the complete Ras signature did not exhibit enrichment for this module. RREB1 is a zinc finger transcription factor ubiquitously expressed in human tissues that binds to RAS-responsive elements (RREs) of gene promoters. Thiagalingam et al. [44] have demonstrated that RREB1 plays a role in Ras and Raf signal transduction in medullary thyroid cancer. In particular, they have shown that the binding of RREB1 to RRE of the calcitonin gene promoter during Ras- or Raf-induced differentiation increases expression of calcitonin in TT human medullary thyroid cancer cells. Our hypothesis is that the 9 genes directly connected to *HRAS* are involved in the downstream signaling of Ras through RREB1. One of them, *EDG4*, has been already found to be correlated to Ras signaling. EDG4 is the receptor for lysophosphatidic acid (LPA), a lipid growth factor and intracellular signaling molecule. It was demonstrated that the expression of a mutated form of Ras GTPase blocked LPA-induced cell migration [35]. This preliminary result suggests that our method is able to enlighten putative regulatory interactions that should be biochemically validated.

4. Conclusions

In the last few years many studies have highlighted the importance of analyzing direct as well as indirect interactions among genes and proteins for unveiling their roles in the onset and progression of complex and multifactorial diseases like tumors. This type of approach is alternative to the classical studies which address the problem of analyzing the association between genes and pathways with the phenotype [1,43]. To this end, many methods have been recently developed to infer gene regulatory networks by using gene expression data [4] in order to reveal putative dependencies among genes and their products. In this paper, we present a comparative study of three different methods to infer networks of conditional dependencies by estimating partial correlation coefficients in the typical situation when the number of observations n is small respect to the number p of variables. The methods and the procedures exploited for their comparison have been developed in the general frameworks of statistical learning theory and regularization theory [46], which constitute state-of-the-art approaches for the analysis and interpretation of data sets composed of a huge number of variables when only a few number of observations is available.

Methods which exploit partial correlation estimates for inferring gene regulatory networks from expression data offer a number of advantages with respect to methods based on mutual information (see for example [33]. In particular, although these methods provide a natural generalization of correlation since they take into account also non-linear dependences between variables, they are not able to assess conditional dependences between two variables in the case the number of conditioning variables is huge as in the context of gene regulatory networks [41].

In our simulation study, we limited our attention to methods which embody an L2 regularization term in their analytical formulation. Such methods, in general, offer more stable solutions with respect to Lasso methods which incorporate L1 regularization terms [27]. The main disadvantage of the adopted techniques for inferring conditional dependency graphs is that they provide non-sparse solutions. To circumvent this problem we have adopted a bootstrap technique which is able to reveal the conditional dependency between two variables with a given statistical significance.

The three analyzed methods were compared through an extensive and biologically inspired simulation study. This choice was adopted because the lack of a validated ground truth relative to biological networks prevents to compare methods by using real gene expression data. In particular, the need of simulated data arises from imperfect knowledge of real networks in cells, from the lack of suitable gene expression datasets, and of control of noise levels. In silico data enable one to check the performance of algorithm against a perfectly known ground truth [4].

Different measures were adopted for assessing the performances of the analyzed methods. Although we did not find a method which consistently outperformed the others in all the carried out simulations, we found that the ℓ_{2C} method provided the most predictive partial correlation estimates, as highlighted by the AUC analysis. More importantly, this method had the highest values of sensitivity showing its ability to infer true conditional dependencies between variables also when a few number of observations is available. Our study has shown that the ℓ_{2C} method is well suited for revealing conditional dependencies when the number of really conditioning variables is small if compared to p as in the case of genomic data.

The application of this method to real biological contexts allowed to infer gene networks with some known regulatory signals. In particular, it revealed a negative significant correlation between the expressions of HMGS and HDS, that we found to be the two hubs in the two isoprenoid pathways in *A. thaliana*.

This means that they respond differently to the several tested experimental conditions and, together with the high connectivity of the two hubs, provides an evidence of cross-talk between genes in the plastidial and the cytosolic pathways. This evidence did not result from studies at level of single gene. Moreover, studies that infer this network by using only low-order partial correlation coefficients find more interactions between the two pathways with respect to the ℓ_{2C} method. A reduced number of edges between the two pathways is plausible considering the different cell compartmentalization of the two isoprenoid biosynthesis pathways.

Moreover, the application of this method to a signature of *HRAS* oncogene permitted to reveal the presence of nine genes connected to *HRAS*, sharing the same Ras-responsive binding site for the transcription factor RREB1. This result suggests that the transcriptional activation of these genes is mediated by a common transcription factor downstream of Ras signaling.

In conclusion, our study has shown that the ℓ_{2C} method is able to infer GRNs with relevant putative interactions and to provide interesting biological hypotheses that should be biochemically validated.

Author's contributions

NA, PFS, SM and FPS conceived the study. PFS, TMC and RA designed the algorithms and conduced the experiments; VCL ana-

lyzed the results from a biological point of view and, together with SM and NA they evaluated and compared the experimental results. All the authors read and approved the final manuscript.

Acknowledgments

We thank Arturo Argentieri for his valuable technical support. This work has been supported in part by Grants from Regione Puglia PO FESR 2007–2013 Progetto BISIMANE (Cod. n. 44), Progetto FIRB RBAP11B2SX, Progetto di Ricerca Finalizzata 2009 RF/2009-1471624.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2013.07.002.

References

[1] Abatangelo L, Maglietta R, Distaso A, D'Addabbo A, Creanza TM, Mukherjee S, et al. Comparative study of gene set enrichment methods. BMC Bioinform 2009;10:275.
[2] Ancona N, Maglietta R, D'Addabbo A, Liuni S, Pesole G. Regularized least squares cancer classifiers from dna microarray data. BMC Bioinform 2005;6:S2.
[3] Anderson TW. An introduction to multivariate statistical analysis. Wiley series in probability and statistics. New York: John Wiley & Sons; 2003.
[4] Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol 2007;3:78.
[5] Bick JA, Lange BM. Metabolic cross talk between cytosolic and plastidial pathways of isoprenoid biosynthesis: unidirectional transport of intermediates across the chloroplast envelope membrane. Arch Biochem Biophys 2003;415:146–54.
[6] Bild A, Yao G, Chang J, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7.
[7] Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 2000;97:12182–6.
[8] Castelo R, Roverato A. A robust procedure for gaussian graphical model search from microarray data with p larger than n. J Mach Learn Res 2006;7:2621–50.
[9] Chang J, Nevins J. Gather: a systems approach to interpreting genomic signatures. Bioinformatics 2006;22:2926–33.
[10] Dempster AP. Covariance selection. Biometrics 1972;28:157–75.
[11] Dikstra RL. Establishing the positive definiteness of the sample covariance matrix. Ann Math Statist 1970;41:2153–4.
[12] Disch A, Hemmerlin A, Bach TJ, Rohmer M. Mevalonate-derived isopentenyl diphosphate is the biosynthetic precursor of ubiquinone prenyl side chain in tobacco by-2 cells. Biochem J 1998;331:615–21.
[13] Dobra A, Hans C, Jones B, Nevins JR, West M. Sparse graphical models for exploring gene expression data. J Multiv Anal 2004;90:196–212.
[14] Edwards D. Introduction to graphical modelling. New York: Springer-Verlag; 1995.
[15] Fawcett T. An introduction to roc analysis. Pattern Recogn Lett 2006;27:861–74.
[16] Fernandez-Medarde A, Santos E. Ras in cancer and developmental disease. Genes Cancer 2011;2:344–58.
[17] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9:432–41.
[18] Friedman J, Hastie T, Tibshirani R. Application of the lasso and grouped lasso to the estimation of sparse graphical models 2010. http://www-stat.stanford.edu/tibs/ftp/ggraph.pdf.
[19] Gilbert H, van der Laan M, Dudoit S. Joint multiple testing procedures for graphical model selection with applications to biological networks. U.C. Berkeley Division of Biostatistics Working Paper Series 245; 2009.
[20] Girosi F, Jones M, Poggio T. Regularization theory and neural networks architectures. Neural Comput 1995;7:219–69.
[21] Harborne JB. Recent advances in the ecological chemistry of plant terpenoids. In: Harborne JB, Tomas-Barberan RA, editors. Ecological chemistry and biochemistry of plant terpenoids. Oxford: Clarendon; 1991.
[22] Hollander M, Wolfe DA. Nonparametric statistical methods. Wiley series in probability and statistics. New York: John Wiley & Sons; 1999.
[23] Kitano H. Systems biology: a brief overview. Science 2002;295:1662–4.
[24] Lange BM, Rujan T, Martin W, Croteau R. Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. Proc Natl Acad Sci USA 2000;97:13172–7.
[25] Laule O, Frholz A, Chang HS, Zhu T, Wang X, Heifetz PB, et al. Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 2003;100:6866–71.
[26] Lauritzen SL. Graphical models. Oxford: Oxford University Press; 1996.
[27] Leng C, Lin Y, Wahba G. A note on lasso and related procedures in model selection. Stat Sinica 2006;16:1273–84.
[28] Li J, Hua X, Haubrock M, Wang J, Wingender E. The architecture of the gene regulatory networks of different tissues. Bioinformatics 2012;28:i509–14.
[29] Lichtenthaler HK, Schwender J, Disch A, Rohmer M. Biosynthesis of isoprenoids in higher plant chloroplasts proceeds via a mevalonate-independent pathway. FEBS Lett 1997;400:271–4.
[30] Luntz A, Brailovsky V. On estimation of characters obtained in statistical procedure of recognition. Techn Kibernet 1969;3.
[31] Madhamshettiwar P, Maetschke S, Davis M, Reverter A, Ragan M. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Med 2012;4:41.
[32] Mansmann U, Jurinovic V. Biological feature validation of estimated gene interaction networks from microarray data: a case study on myc in lymphomas. Brief Bioinform 2011;12:230–44.
[33] Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, et al. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform 2006;7:S7.
[34] Meinshausen N, Buhlmann P. High-dimensional graphs and variable selection with the lasso. Ann Stat 2006;34:1436–62.
[35] Park S, Schinkmann K, Avraham S. Raftk/pyk2 mediates lpa-induced pc12 cell migration. Cell Signal 2006;18:1063–71.
[36] Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. J Am Stat Assoc 2009;104:735–46.
[37] Raudys S, Duin RPW. Expected classification error of the fisher linear classifier with pseudoinverse covariance matrix. Pattern Recogn Lett 1998;19:385–92.
[38] Rodriguez-Concepcion M, Fores O, Martinez-Garcia JF, Gonzalez V, Phillips M, Ferrer A, et al. Distinct light-mediated pathways regulate the biosynthesis and exchange of isoprenoid precursors during arabidopsis seedling development. Plant Cell 2004;16:144–56.
[39] Schäfer J, Strimmer K. An empirical bayes approach to inferring large-scale gene association networks. Bioinformatics 2005;21:754–64.
[40] Simon N, Tibshirani R. A permutation approach to testing interactions in many dimensions. arXiv:1206.6519 [stat.ML].
[41] Soranzo N, Bianconi G, Altafini C. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. Bioinformatics 2007;23:13:1640–7.
[42] Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003;302:249–55.
[43] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 2005;102:15545–50.
[44] Thiagalingam A, De Bustros A, Borges M, Jasti R, Compton D, Diamond L, et al. is involved in the differentiation response to ras in human medullary thyroid carcinomas. Mol Cell Biol 1996;16:5335–53045.
[45] Toh H, Horimoto K. System for automatically inferring a genetic network from expression profiles. J Biol Phys 2002;28:449–64.
[46] Vapnik V. Statistical learning theory. New York: John Wiley & Sons; 1998.
[47] Whittaker J. Graphical models in applied multivariate statistics. New York: John Wiley & Sons; 1990.
[48] Wille A, Buhlmann P. Sparse graphical gaussian modelling of the isoprenoid network in *Arabidopsis thaliana*. Genome Biol 2004;5:R92.
[49] Witten DM, Tibshirani R. Covariance-regularized regression and classification for high dimensional problems. J R Stat Soc B 2009;71:615–36.
[50] Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. Biometrika 2007;94:19–35.
[51] Zhang B, Li H, Riggins R, Zhan M, Xuan J, Zhang Z, et al. Differential dependency network analysis to identify condition-specific topological changes in biological networks. Bioinformatics 2009;4:526–32.